

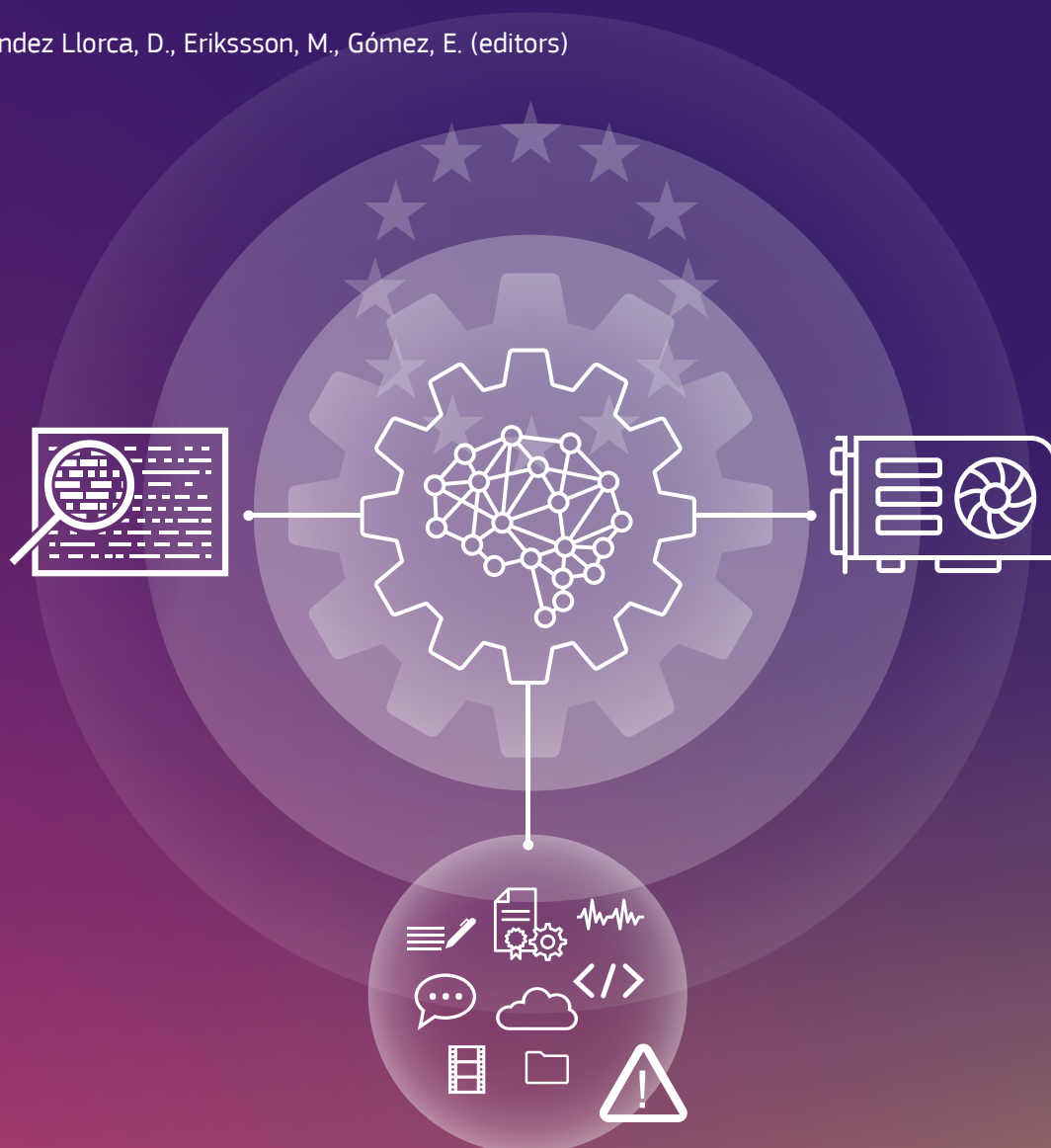
# A Proposal to Identify High-Impact Capabilities in General-Purpose AI Models

Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act

Hobbhahn, M., Hovy, D., Vanschoren, J.

Fernández Llorca, D., Eriksson, M., Gómez, E. (editors)

**2025**



This publication is an External Study report prepared for the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

### Contact Information

Name: David Fernández Llorca

Address: European Commission, Joint Research Centre (JRC) Edificio Expo, c/Inca Garcilaso, 3, 41092 Seville - Spain

Email: david.fernandez-llorca@ec.europa.eu

### The Joint Research Centre: EU Science Hub

<https://joint-research-centre.ec.europa.eu>

JRC143258

PDF ISBN 978-92-68-31572-9 doi:10.2760/8206407 KJ-01-25-469-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

How to cite this report: Hobbhahn, M., Hovy, D., Vanschoren, J. *A Proposal to Identify High-Impact Capabilities in General-Purpose AI Models*, Fernández Llorca, D., Eriksson, M., Gómez, E. (editors), Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/8206407>, JRC143258.

# Contents

Abstract.....	2
Acknowledgements.....	3
Note from the Editors.....	4
Executive summary.....	6
1 Introduction.....	7
1.1 Benefits of the PCA approach.....	7
1.2 Limitations.....	7
2 Benchmark selection criteria.....	9
3 Specific selection of benchmarks.....	11
3.1 MMLU-Pro.....	11
3.2 GPQA-diamond.....	11
3.3 MATH-level-5.....	11
3.4 HumanEval.....	12
3.5 SWE-Bench-verified (a subset).....	12
3.6 MLE-Bench (a subset).....	12
4 Benchmark-score aggregation.....	13
4.1 Simplified approach.....	14
4.2 Tiered approach.....	14
5 Additional Considerations.....	16
5.1 Detailed procedure for measurements.....	16
5.2 6-months updates.....	17
5.3 Mitigation measures to prevent gaming & other actions by GPAI model providers.....	17
6 Conclusions.....	19
References.....	20
List of abbreviations and definitions.....	22
List of figures.....	23
List of tables.....	24
Annexes.....	25
Annex 1. Concrete example of benchmark aggregation.....	25
Annex 2. Robustness tests.....	29

## **Abstract**

This report proposes a scientific methodology to identify high-impact capabilities in general-purpose AI (GPAI) models, defined in the EU AI Act as capabilities of the most advanced GPAI models. High-impact capabilities play an important role in the EU AI Act since GPAI models with high-impact capabilities are classified as GPAI models with systemic risk. The approach presented by this report is based on observational scaling laws using Principal Components Analysis (PCA) from a set of existing benchmarks, allowing for the extraction of a low-dimensional capability measure that can be used to identify models with high-impact capabilities through setting a suitable threshold. The proposed method involves selecting a diverse set of benchmarks that measure general capabilities, such as MMLU-Pro, GPQA-diamond, MATH-level-5, and HumanEval, and aggregating their scores using a weighted threshold-based metric. The weights are determined by the PCA approach, and the threshold is based on a reference model, to be set by the enforcement authority based on legal, policy, and risk considerations. The report also discusses additional considerations, including proposing for a multi-disciplinary expert group to oversee benchmark selection, and for the approach to be updated every 6 months to account for rapid developments in AI, and suggesting mitigation measures to prevent companies from strategically underperforming on benchmarks. By providing a practical and robust way to assess high-impact capabilities, this methodology aims to contribute to the development of a more comprehensive approach to evaluating GPAI models.

## **Acknowledgements**

The editors would like to thank the JRC colleagues who have helped us develop the work resulting in this collection, the AI Office for its inputs, as well as those who have kindly agreed to review the drafts and the final external study reports.

### ***Authors***

Hobbhahn, Marius  
Hovy, Dirk  
Vanschoren, Joaquin

### ***Editors***

Fernández Llorca, David  
Eriksson, Maria  
Gómez, Emilia

## Note from the Editors

The EU AI Act entered into force on 1 August 2024, with the aim of promoting innovation in and uptake of AI in the Union, while ensuring a high level of protection of health, safety and fundamental rights, including democracy and the rule of law. Chapter V of the AI Act outlines obligations for the providers of general-purpose AI (GPAI) models, which are AI models *"trained with a large amount of data using self-supervision at scale, that [display] significant generality and [are] capable of competently performing a wide range of distinct tasks [...] and that can be integrated into a variety of downstream systems or applications"*. Moreover, the chapter specifies additional obligations for the most advanced GPAI models, those that pose systemic risks, which are classified as such according to criteria established in Article 51 and Annex XIII. From 2 August 2025, the obligations for providers of GPAI models and GPAI models with systemic risk enter into application.

The European Commission's Joint Research Centre (JRC) has been providing scientific support throughout the legislative process of the AI Act since 2020. After the Council and Parliament reached a final agreement in December 2023, the JRC initiated an internal study, which included two external experts, focusing on the technical aspects of Chapter V that likely required further clarification. This study generated an internal report titled "General Purpose AI Models under the AI Act" (Hernández-Orallo et al., 2024), which provided preliminary insights into compute, generality, capabilities, and systemic risks. One of the clear conclusions of this preliminary study was that further scientific work was necessary.

Between September 2024 and June 2025, the JRC setup and managed, in close collaboration with the EU AI Office, a pool of 15 external experts with diverse expertise and backgrounds. This expert pool produced further technical scientific input on key aspects of Chapter V, through the development of methodologies for categorising AI models as GPAI models and for classifying GPAI models as GPAI models with systemic risk, to inform implementation of the EU AI Act. The experts also provided input on the recently published Commission guidelines on the scope of obligations for providers of GPAI models (European Commission, 2025), as part of the public multi-stakeholder consultation. The primary outcome of this expert pool is this **Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act**, which comprises a total of six external scientific study reports.

Although more documents may be added to the collection in the future, as of the writing of this editorial, the titles of the external reports included in the collection are:

- Training Compute Thresholds - Key Considerations for the EU AI Act
- A Framework for General-Purpose AI Model Categorisation
- A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes
- A Proposal to Identify High Impact Capabilities of General-Purpose AI Models
- The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Model
- General-Purpose AI Model Reach as Criterion for Systemic Risk

The overall objective of this collection is twofold. On the one hand, it aims to contribute to broadening the understanding and discussion of the technical and scientific issues related to GPAI models and the identification of systemic risks. On the other hand, it seeks to provide a solid scientific basis for informing the implementation of Chapter V of the EU AI Act, which has recently entered into application. It is clear that we are dealing with complex issues, where a clear scientific consensus has yet to be established, and which require a certain degree of flexibility. Nevertheless, this is part of the

necessary effort to promote innovation and the uptake of AI, while ensuring protection for human health, safety, and fundamental rights in Europe.

These external scientific studies cover aspects regarding the presumption of having high impact capabilities based on cumulative amount of computation used for training (Article 51(2)), notification conditions (Article 52), the definition of a GPAI model (Article 3(63)), and considerations for GPAI models being classified as GPAI models with systemic risk based on capability benchmarks, safety benchmarks and reach (Article 51(1) and Annex XIII).

These studies reflect the outcome of the scientific and technical analysis of a series of external experts to the Commission. In some cases, they present a state-of-the-art review, while in others, they propose methodologies based on solid scientific evidence, while acknowledging significant uncertainty, as many of these problems still lack a widely accepted solution. The content, analysis, recommendations, and suggestions should then not be interpreted in any way as the position of the Commission, nor of the JRC editors in particular, but rather as the opinion of the authors.

## Executive summary

We propose a method for assessing whether a general-purpose AI (GPAI) model has high-impact capabilities, using a technique based on observational scaling laws. The technique uses Principal Component Analysis (PCA) (Ruan et al., 2024) from results on a basket of benchmarks already existing today.

The selection of benchmarks to be included could be overseen by a multi-disciplinary expert group. We provide recommendations and guidelines for this selection.

Currently, this "basket of benchmarks" to assess whether or not a GPAI model has high-impact capabilities could include MMLU-Pro, GPQA-diamond, MATH-level-5, HumanEval, SWE-Bench-verified (subset), and MLE-Bench-lite. We include a concrete illustrative example using only the first four benchmarks and 30+ models.

We use a weighted threshold-based metric where the weights come from the PCA approach, and the threshold is based on a reference model. Although the reference model should be set by the enforcement authority based on legal, policy and risk considerations, for the purpose of this report we have taken GPT-4o as the reference model. Nevertheless, the method proposed on this report works for any choice of reference model, the threshold simply needs to be updated if the reference model is changed. We provide the aggregation equation found in our example below.

To prevent every GPAI model having to be tested on all of these benchmarks, we use a tiered approach where models are not considered to have high-impact capabilities if they perform poorly on individual benchmarks.

We provide a list of mitigation measures to prevent companies actively trying to underperform (sandbag) these benchmark scores.

If the decision is made to make the assessment methodology, choice of benchmarks and reference model public in order to allow providers to self-assess whether their GPAI model has high-impact capabilities, a detailed description about how to run these benchmarks should be provided to offer a standardised procedure to providers.

We recommend that the set of benchmarks, thresholds, models, and measurement protocols be updated every six months to allow for flexibility in the face of rapid developments in AI. This could be done with the help of an expert group, and might include private benchmarks to avoid the risk of benchmarks contamination.

# 1 Introduction

There are empirical compute scaling laws for LLMs (Hoffmann et al., 2022). The key finding is that performance follows a power-law relationship with training compute. This is not a strict law, but rather a solid empirical trend. Models with higher scaling laws are more likely to have emergent capabilities when trained for longer, and these emergent capabilities may carry systemic risk (e.g. novel reasoning, deception, or strategic behaviour).

We note that predicting model capability from scaling laws is still elusive (Schaeffer et al., 2024) and should be part of a more comprehensive approach. At the same time, it is potentially a more robust approach than using FLOP or model size as metrics. For instance, model capability can depend greatly on the quality of the training data used, which is captured in scaling laws but not in training compute (FLOP).

A paper on predicting LLM performance (Ruan et al., 2024) shows that model performance on benchmark datasets can be reliably used to 'find scaling laws' (or at least some empirical relationship between model performance and compute). It shows that you can apply PCA to the benchmark results and 97% of the variance in capabilities is explained by the top-3 components. Hence, we can extract a low-dimensional 'capability measure' that demonstrates a (log-linear) relationship with compute scale measures. In the paper, the top-3 components correlate with benchmarks on general capabilities (MMLU (Hendrycks et al., 2020)), reasoning (GSM8K (Cobbe et al., 2021) and HellaSwag (Zellers et al., 2019)) and programming (HumanEval (Chen et al., 2021)).

## 1.1 Benefits of the PCA approach

**The approach is easily applicable.** We can run GPAI models on benchmarks to predict whether a model may have high-impact capabilities, once a reference has been set for what should count as the minimum capabilities required to have 'high-impact capabilities'. We believe this is one of the most practical ways to assess whether or not a model has high-impact capabilities. It only requires running the model on benchmarks, which is simple (and model developers typically do this anyway).

**The results are robust.** ML benchmarks have repeatedly shown to have 'external validity', i.e., the relative performance rankings of ML algorithms generalise across different benchmarks (Salaudeen and Hardt, 2024). This aligns with the core principle of the PCA approach: a low-dimensional capability space exists for LLMs independent of the exact benchmarks used. In other words, scaling laws could (in theory) be formulated in terms of latent capability values that we can measure through benchmarks, and these values seem to be fairly independent from the benchmarks used.

However, the benchmarks should still be selected with care<sup>1</sup>, taking into account at least the following considerations:

1. They should measure capabilities that we care about (i.e., measures of general capabilities aligned with the EU AI Act definitions).
2. To arrive at meaningful latent capability values, we should select benchmarks which are as diverse as possible.

## 1.2 Limitations

There are a number of limitations to this technique that should be considered:

---

<sup>1</sup>See (Eriksson et al., 2025) for an interdisciplinary review of current issues with AI benchmarks.

- The proposed approach is solely focused on evaluating whether a GPAI model has high-impact capabilities, understood as the “capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models” (Article 3(64)). However, although having high-impact capabilities is one of the conditions for a GPAI model to be classified as a GPAI model with systemic risk (Article 51(a) of the AI Act), the consideration of the proposed approach as a suitable technical tool or methodology for this purpose is outside the scope of this report.
- The interpretability of the PCA-based approach has not been fully demonstrated, and a more in-depth ablation analysis would be necessary to address this limitation.
- The generality of the method is inherently tied to the generality of the models and benchmarks employed. To ensure a comprehensive evaluation, it is essential to select a diverse and appropriate set of benchmarks that accurately represent the capability space of interest. For instance, relying solely on mathematical benchmarks would be insufficient, as it would not adequately capture the full range of capabilities. Similarly, including coding benchmarks in an assessment of 'bio' capabilities would not be effective.
- Previous factor analysis studies (see (Burnell et al., 2023) and (Ilić and Gignac, 2024)) have consistently identified 1-3 dominant factors when extracting latent factors or components that explain the variance in the data. However, these "populations" of models and benchmarks are volatile, exhibiting complex structural dependencies that can impact the stability and interpretability of the results.

Given the above limitations, we recommend that any policy that makes use of the approach proposed in this report should be flexible enough to allow the evaluations to be replaced by better assessment techniques as they become available.

## 2 Benchmark selection criteria

This section suggests criteria that could guide the selection of benchmarks that underpins the method presented in this report for assessing high-impact capabilities. In addition, we recommend for initiatives that stimulate the development of better benchmarks to be developed. For example, a multi-disciplinary benchmarking expert group could be set up and called upon for recommendations on which benchmarks could be included, or even for the creation of new benchmarks where needed.

### **The benchmark should be generally accepted:**

- Simple indicators could include the number of paper citations (e.g. more than 100), whether the benchmark is generally used in frontier model comparisons by the companies themselves, or whether it measures general capabilities rather than, for example, specific failure modes or propensities.
- Notably, these indicators can be arbitrary and prone to cherry-picking, and therefore any entity responsible for selecting benchmarks should have the final say in whether this criterion is needed and when it is met.

### **Generality:**

- Does the benchmark measure general capabilities, or does it make the benchmark set more diverse as a whole (i.e., it adds skills not tested by other benchmarks). In other words, does adding benchmark X add coverage of capabilities would otherwise not be provided.

### **The benchmark cannot be saturated:**

- A benchmark where the three most capable GPAI models are within 1 percentage point of the natural ceiling of the benchmark can be counted as saturated. The range is from the natural lower bound to the natural upper bound, e.g., 0% to 100% for accuracy scores but can also be applied to other metrics as long as they have a ceiling. It is possible that the "effective ceiling" of a benchmark is lower, e.g. because some data points are mislabelled. We would recommend that the decision to lower the ceiling lies in the discretion of the regulator.

### **The benchmark has to be publicly available:**

- Public benchmarks are particularly important if providers are expected to run the benchmarks themselves.
- All datapoints required to run the benchmark should be publicly available, if providers are expected to self-assess their models. In this case, the datapoints should be available, not just through an API. It would be desirable for each provider to have the option to establish their own benchmarking procedure.
- The benchmark should be available in a format that is easy to use for AI developers, e.g. in a commonly used dataformat or hosted in a commonly used repository. We would recommend that the competent authority has discretion about deciding which datasets are easy to use.
- The enforcement authority may decide to use or provide "hidden test sets" for benchmarks, e.g. to test for overfitting or sandbagging.
- To avoid risks of benchmark contamination, the entity responsible for selecting benchmarks could consider setting up an API with hidden benchmarks.

**The benchmark cannot be too expensive to run:**

- We recommend that the entity responsible for selecting benchmarks should have the discretion to determine what constitutes a benchmark that is "too expensive to run".
- Heuristics: does it cost more than \$1K to run the benchmark on a public API for a model of similar quality? Does it take a single skilled Full-Time Equivalent (FTE) more than 3 days to set up and run the benchmark?
- In case the benchmark is too expensive to run, we recommend that the entity responsible for selecting benchmarks specifies an exact subset of datapoints for the benchmark such that it is not too expensive to run.

### 3 Specific selection of benchmarks

In the following, we identify a set of public benchmarks that are proposed as one possible basket of benchmarks for assessing high-impact capabilities.

#### 3.1 MMLU-Pro

- MMLU (Hendrycks et al., 2020) is the canonical LLM capability benchmark, thus clearly meeting the criterion of being generally known. It is publicly available and not too expensive to run.
- MMLU has a couple of potential problems, such as incorrect labels and only four options per question. Thus, we recommend using MMLU-pro (Wang et al., 2024), which has 10 answer options and improved labels.
- It is possible that MMLU-pro is almost saturated. While we think that it might not take long for MMLU-pro to be saturated, most of the best performances come from extensive prompt setups with approaches like best-of-8 with 32-shot. We expect that under simpler conditions, e.g. CoT & 0-shot, the benchmark is not yet saturated in a meaningful way.
- The MMLU-Pro paper has 516 citations and the original MMLU paper has 4828 as of July 2025.
- Link to accessible implementation:  
[https://github.com/UKGovernmentBEIS/inspect\\_evals/tree/main/src/inspect\\_evals/mmlu\\_pro](https://github.com/UKGovernmentBEIS/inspect_evals/tree/main/src/inspect_evals/mmlu_pro).

#### 3.2 GPQA-diamond

- GPQA diamond is a benchmark specifically designed to contain hard questions that can take a human expert more than 30 minutes to solve (Rein et al., 2024). It is commonly used as the canonical "hard benchmark", is publicly available and not too expensive to run.
- It is clearly not saturated as of February 2025. Though some of the questions might be unanswerable or the labels might be incorrect.
- The paper has 823 citations as of July 2025.
- Link to accessible implementation:  
[https://github.com/UKGovernmentBEIS/inspect\\_evals/tree/main/src/inspect\\_evals/gpqa](https://github.com/UKGovernmentBEIS/inspect_evals/tree/main/src/inspect_evals/gpqa)

#### 3.3 MATH-level-5

- MATH is a general reasoning benchmark with 5 levels. It is publicly available and not too expensive to run.
- While MATH level 1-4 are likely saturated, level 5 does not seem to be (February 2025).
- The paper has 2411 citations as of July 2025.
- Link to accessible implementation:  
[https://github.com/UKGovernmentBEIS/inspect\\_evals/tree/main/src/inspect\\_evals/mathematics](https://github.com/UKGovernmentBEIS/inspect_evals/tree/main/src/inspect_evals/mathematics)

### 3.4 HumanEval

- The most common "simple but not saturated" coding benchmark (Chen et al., 2021). It is publicly available and not too expensive to run.
- It is plausible that HumanEval, even with 0-shot and CoT is saturated and thus does not meaningfully differentiate between frontier models any more as of May 2025.
- The paper has 5561 citations as of July 2025.
- Link to accessible implementation:  
[https://github.com/UKGovernmentBEIS/inspect\\_evals/tree/main/src/inspect\\_evals/humaneval](https://github.com/UKGovernmentBEIS/inspect_evals/tree/main/src/inspect_evals/humaneval).

### 3.5 SWE-Bench-verified (a subset)

- Our basket of benchmarks should include at least one, but likely multiple, benchmarks that test the agentic capabilities of models. SWE-Bench-verified is the go-to example for coding agent capabilities (OpenAI, 2024a).
- It is plausible that running all 500 examples is too expensive. Thus, we suggest the entity responsible for selecting the benchmarks picks a list of 50 representative samples of the benchmark and then only requires running this smaller subset as SWE-Bench-verified-mini.
- We made a subset for this purpose here:
  - <https://github.com/mariushobbhahn/SWEBench-verified-mini>.
- Link to accessible implementations:
  - [https://huggingface.co/datasets/princeton-nlp/SWE-bench\\_Verified](https://huggingface.co/datasets/princeton-nlp/SWE-bench_Verified).
  - [https://github.com/UKGovernmentBEIS/inspect\\_evals/tree/main/src/inspect\\_evals/swe\\_bench](https://github.com/UKGovernmentBEIS/inspect_evals/tree/main/src/inspect_evals/swe_bench).

### 3.6 MLE-Bench (a subset)

- The same reasoning as for SWEBench applies.
- The authors of MLE-Bench (OpenAI, 2024b) made a "lite" split that we would recommend for this purpose:
  - <https://x.com/junshernchan/status/1876315258819944792>
- Link to accessible implementation: <https://github.com/openai/mle-bench/>

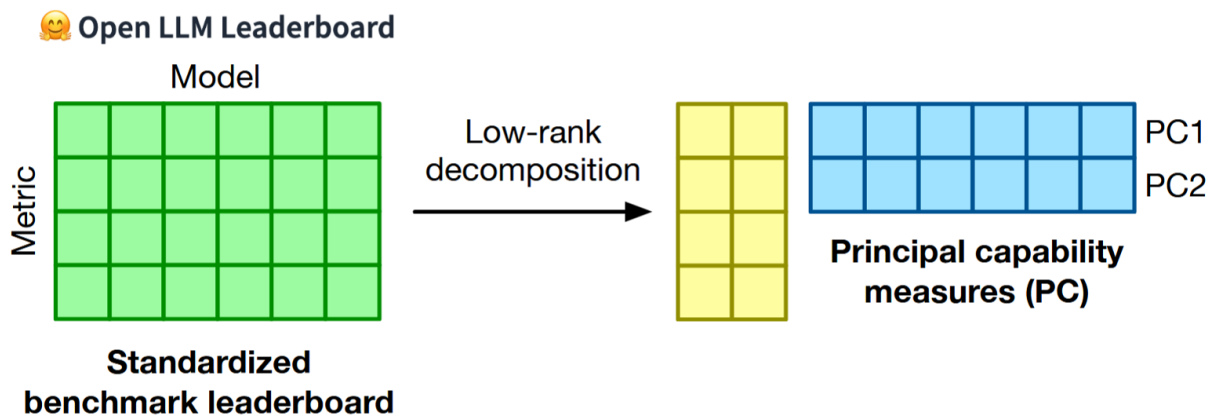
## 4 Benchmark-score aggregation

To determine whether a model has high-impact capabilities from a set of benchmark results, we propose using a single weighted average score that combines all individual benchmark scores.

The weight for the weighted average are determined using observational scaling laws (Ruan et al., 2024). We provide a significantly more detailed justification for why a PCA-based approach captures the core component of what we are looking for in Annex 1. To determine the weights, the regulator would need to keep a table T of models and benchmark scores that span all benchmarks and a large range of model performances.

Then, the PCA of table T as described in the observational scaling laws paper can be computed.

**Figure 1:** Illustration of the proposed computation of the principal capability measures.



Source: Edited from (Ruan et al., 2024).

The first principal component, PC-1, can be interpreted as the component describing "general capabilities", which captures what we are aiming for with the approach. We standardise the columns before computing the PCA because scores could come from any task and domain and do not have to be between 0 and 1.

Concretely, here are the exact steps we pursue:

1. We compute the mean  $m$  and scale factor  $s$  for all benchmarks (columns) individually.
2. We compute the PCA of the standardised matrix. Then we take the projection vector for  $PC1$ , which we call  $PC1_{component}$ .
3. For any new datapoint (i.e. model), we compute the  $PC1$  score by:

(a)  $Score = \sum_i pc1_{component,i} \times (x_i - m_i) / s_i$

(b) However, for fixed mean  $m$ ,  $PC1_{component}$  and scale  $s$ , this can be simplified to:

i.  $weights = PC1_{component} / s$

ii.  $bias = - \sum_i m_i \times pc1_{component,i} / s_i$

iii.  $Score = (\sum_i x_i \times weights_i) + bias$

In our particular case, using the first four benchmarks, this results in (numbers only displayed to 2 sig. digits for visibility):

$$Score = 2.60 \times \text{MMLU-Pro} + 3.27 \times \text{GPQA-diamond} + 2.30 \times \text{MATH-level-5} + 1.98 \times \text{HumanEval} - 5.25$$

Furthermore, we set the threshold for inclusion based on the scores of a model that the regulator would consider on the border of being included or not. For example, the threshold value determined by GPT-4o is: 0.99 (this is by accident, 0.99 does not have any special meaning here).

For a practical example, see Annex 1. We test the robustness of the approach to removing benchmarks and models from the training set in Annex 2. We find that the approach is clearly sufficiently robust for all practical intents and purposes.

## 4.1 Simplified approach

As an alternative to the above approach, we can consider a simplified version that works as follows:

1. We normalise each benchmark such that it has a range from 0 to 1.
2. We run a PCA-based on the normalised column values. Then we either:
  - (a) Compute the weights for each benchmark as the normalised contribution to the first PCA component (which happens to be almost perfectly uniformly distributed in our example case, e.g., 0.25 for each of the four benchmarks)
  - (b) OR the enforcement authority sets the weights based on their own estimates of the importance of these benchmarks (potentially based on expert consensus).
3. The final score is then also between 0 and 1 and might thus be more interpretable than the PCA-based score.
4. The threshold would be chosen through reference to an existing model as well, e.g. OpenAI's o1 or Claude Sonnet 4.

We think this approach is simpler to understand to a large audience and potentially loses only very little fidelity. Thus, if the enforcement authority wants to provide the simplest and yet most reasonable approach possible, we would recommend this simplified version.

## 4.2 Tiered approach

Running all of these benchmarks is unnecessarily expensive, especially when a GPAI model is clearly not capable enough to be considered a GPAI model with high-impact capabilities. Therefore, we suggest a tiered approach that can be implemented as follows (note that the percentages are simply illustrative and should be adapted as appropriate based on technical, policy and legal considerations):

- In all cases, the AI model has to be run using the procedure in the section on exact procedures for measurements.
- **Step 1:** Run the model on MMLU-Pro. If the model achieves a score lower than 65% it is not assumed to have high-impact capabilities.
  - Most model providers run their models on MMLU-Pro anyway, therefore it should not impose an additional cost to them.
- **Step 2:** Run the model on GPQA-diamond and MATH-level-5. If the model achieves a score lower than 40% and 45% respectively, then it is not assumed to have high-impact capabilities.
  - These are cheap evaluations to run. Therefore, it should not pose a big burden on the model provider.
- **Step 3:** Run the model on the remaining benchmarks and compute the overall score as described in the aggregation section.
  - Subsets of SWE-Bench-verified and MLE-Bench are likely the most resource intensive. Therefore, we only ask model providers to measure it in the final step.

## 5 Additional Considerations

### 5.1 Detailed procedure for measurements

In the case where providers are expected to run the benchmarks themselves, we recommend that the regulator provides exact descriptions of how to run the benchmarks that are specified to prevent either over- or underperformance by models and standardise performance across model providers.

We recommend that the model provider is not allowed to actively train on the test set of these benchmarks, neither to increase nor to decrease performance. These effects could be identified either through statistical analysis or through reports from whistleblowers.

#### Q&A benchmarks:

1. MMLU-Pro: CoT & 0-shot, temperature=1, one run across the entire benchmark.
2. GPQA-diamond: CoT & 0-shot, temperature=1, average of 3 runs across the entire benchmark.
3. MATH-level-5: CoT & 0-shot, temperature=1, average of 3 runs across the entire benchmark.
4. HumanEval: CoT & 0-shot, temperature=1, average of 3 runs across the entire benchmark.

If appropriate, the regulator could provide a concrete implementation of all benchmarks with exact specifications implemented in a single framework, such as Inspect. Since Inspect already covers all of the suggested benchmarks, the necessary adaptations would be minimal.

#### Agent benchmarks:

We have to find a recommendation that is likely:

1. Reasonably expectable for all developers.
2. Not too specific to the benchmark.
3. Not terribly underelicited.

Given that we merely need to find *relative capability differences* instead of eliciting maximal performance, we do not require the best possible scaffolding.

A suggestion would be that the scaffolding needs to include:

1. A bash tool.
2. A python tool.
3. An edit tool that enables more targeted code edits to improve coding ability, e.g. similar to the tool used by AIDER(Aider AI, 2025).

In general, these specifications should be updated to broadly align with the trends in agentic scaffolding. They can be conservative, but they should not lag too far behind the state of the art.

If appropriate, the regulator could provide an implementation of both the agent and the benchmark in a single framework, such as Inspect, offering a straightforward option for deployers. Inspect already supports SWE-Bench (and possibly MLE-Bench soon). However, the Inspect "basic\_agent" lacks robust editing tools, and the exact subset of the respective benchmarks should be specified (e.g., SWE-Bench-mini and MLE-Bench-lite).

## 5.2 6-months updates

Based on the current pace of progress in the field, we recommend updating the following aspects of the proposed approach every 6 months:

- Benchmarks: saturated benchmarks have to be phased out and potentially new relevant benchmarks have to be phased in.
- Thresholds: thresholds have to be updated based on trends in capabilities.
- Models: the update of the thresholds will obviously impact the number of GPAI models considered as having high-impact capabilities.
- Measurement descriptions: descriptions of how to run each benchmark, if appropriate.

If appropriate, the regulator could provide updated implementations in an open-source framework, such as Inspect, for all updated benchmarks and specifications.

### Benchmarks

We recommend that saturated benchmarks (as defined in the "Benchmark selection" section) should be phased out of the benchmark basket. Benchmarks deemed important could then be phased in, provided they meet the criteria outlined in the "Benchmark selection" section.

### Thresholds

The benchmark thresholds should be adapted to match the evolving collection of the "most advanced models" (recalling that high-impact capabilities are capabilities of the most advanced models). That is, thresholds should be updated as necessary so that the least capable model among those considered to be "most advanced", which is classified using our proposed methodology as a GPAI model with high-impact capabilities, is distinguished from the most capable model that is not considered to be part of the most advanced group.

### Models

The threshold updates will consequently affect some models that may no longer be considered to have high-impact capabilities if they fall below the updated, increased threshold. In theory, the opposite effect is also possible if the threshold is lowered, i.e. models that were previously not considered to have high-impact capabilities may subsequently be considered as having such capabilities. However, in practice, decreasing the threshold to identify the capabilities of the most advanced GPAI models is counter-intuitive, given that model capabilities are continually improving.

### Measurement descriptions

In the event that the regulator needs to modify the measurement descriptions, this information should be made publicly available, if providers are expected to run the benchmarks themselves. We can expect that, in most cases, these updates will pertain to changes in how to measure agentic capabilities, as this is a more emerging field compared to static benchmarking. If possible, we would recommend that they provide implementations of the new suggestions, such as agents and benchmarks, in an open-source framework such as Inspect.

## 5.3 Mitigation measures to prevent gaming & other actions by GPAI model providers

There are a couple of strategies that an adversarial actor could attempt:

1. **Strategic underperformance**, i.e. they might try to make their model less competent on the relevant benchmarks. For example:
  - (a) The model developer might strategically try to make their model less capable across all benchmarks.
  - (b) The model developer might strategically try to make their model less capable on a particular evaluation. Presumably, they would want to underperform on one that reduces the overall score the most.
2. **Strategic underelicitation**
  - (a) The model developer might try to intentionally underelicit the results even if the model is theoretically capable of achieving them.

Possible responses:

1. **Strategic underperformance**
  - (a) If the model developer trains the model to have low performance on the respective benchmarks, they also reduce their public reputation because it implies that their model is less capable.
  - (b) Model providers should not be allowed to actively train on the test set of these benchmarks, regardless of whether it is to increase or decrease performance. This behaviour can be detected through statistical analysis or whistleblower reports.
2. **Strategic underelicitation**
  - (a) Underelicitation seems unlikely in the Q&A benchmarks since the proposal specifies the procedure to run it in a lot of detail. For agent evaluations, it might be easier to strategically underelicit.
  - (b) There may be natural reputational deterrents to underelicitation. Since these benchmarks are conventionally used to demonstrate model capabilities, a low score may be perceived as indicative of a poorly performing model by the customers of the provider.
  - (c) If the regulator is able to elicit substantially better performance with the same methods the model provider has claimed they were using, then the performance elicited by the authority should take precedent over that elicited by the provider.

## 6 Conclusions

In this report, we provide a scientific methodology to identify high-impact capabilities in general-purpose AI (GPAI) models, as defined in the EU AI Act, while minimising the burden on model providers.

The proposed approach is based on observational scaling laws using Principal Component Analysis (PCA) from a set of existing benchmarks, allowing for the extraction of a low-dimensional capability measure.

We demonstrate that this method can be used to assess whether a GPAI model has high-impact capabilities, given a reference model, and we provide a concrete example using a subset of benchmarks, including MMLU-Pro, GPQA-diamond, MATH-level-5, and HumanEval.

We propose the implementation of a tiered approach, as well as mitigation measures to prevent gaming and underelicitation to ensure the integrity and effectiveness of the methodology.

The approach is designed to be flexible and adaptable, with regular updates to the benchmarks, thresholds, and measurement protocols recommended every 6 months to keep pace with the rapid advancements in AI.

By providing a practical and robust way to assess high-impact capabilities, this methodology aims to contribute to the development of a more comprehensive approach to evaluating GPAI models, ultimately supporting the goals of the EU AI Act and promoting a safer and more responsible AI ecosystem.

## References

- Aider AI, 'AI Pair Programming in Your Terminal'. <https://github.com/Aider-AI/aider>, 2025. Accessed: 2025-07-17.
- Burnell, R., Hao, H., Conway, A. R. and Orallo, J. H., 'Revealing the structure of language model capabilities', [arXiv preprint arXiv:2306.10062](https://arxiv.org/abs/2306.10062), 2023.
- Chen, M. et al., 'Evaluating large language models trained on code', [arXiv preprint arXiv:2107.03374](https://arxiv.org/abs/2107.03374), 2021.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R. et al., 'Training verifiers to solve math word problems', [arXiv preprint arXiv:2110.14168](https://arxiv.org/abs/2110.14168), 2021.
- Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gómez, E. and Fernández-Llorca, D., 'Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation', [Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society \(AIES-25\)](https://proceedings.aai.acm.org/ai-ethics-and-society-2025), 2025.
- European Commission, 'Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)'. <https://ec.europa.eu/newsroom/dae/redirection/document/118340>, 2025.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J., 'Measuring massive multitask language understanding', [arXiv preprint arXiv:2009.03300](https://arxiv.org/abs/2009.03300), 2020.
- Hernández-Orallo, J., Sevilla, J., Gómez, E. and Fernández-Llorca, D., 'General-Purpose AI Models in the AI Act: Preliminary Insights on Capabilities, Generality, Systemic Risks and Compute', [European Commission, Joint Research Centre, JRC139341](https://www.ec.europa.eu/commission/press-materials/press-conferences-events/joint-research-centre-2024_en), 2024.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A. et al., 'Training compute-optimal large language models', [arXiv preprint arXiv:2203.15556](https://arxiv.org/abs/2203.15556), 2022.
- Ilić, D. and Gignac, G. E., 'Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement?', [Intelligence](https://doi.org/10.1080/10888422.2024.2311158), Vol. 106, 2024, p. 101858.
- OpenAI, 'Introducing SWE-bench Verified'. <https://openai.com/index/introducing-swe-bench-verified/>, 2024a. Accessed: 2025-07-17.
- OpenAI, 'MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering'. <https://openai.com/index/mle-bench/>, 2024b. Accessed: 2025-07-17.
- Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J. and Bowman, S. R., 'Gpqa: A graduate-level google-proof q&a benchmark', In 'First Conference on Language Modeling', .
- Ruan, Y., Maddison, C. J. and Hashimoto, T. B., 'Observational scaling laws and the predictability of language model performance', [Advances in Neural Information Processing Systems](https://arxiv.org/abs/2406.04391), Vol. 37, 2024, pp. 15841–15892.
- Salaudeen, O. and Hardt, M., 'Imagenot: A contrast with imagenet preserves model rankings', [arXiv preprint arXiv:2404.02112](https://arxiv.org/abs/2404.02112), 2024.
- Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S. and Koyejo, S., 'Why has predicting downstream capabilities of frontier ai models with scale remained elusive?', [arXiv preprint arXiv:2406.04391](https://arxiv.org/abs/2406.04391), 2024.
- Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z. et al., 'Mmlu-pro: A more robust and challenging multi-task language understanding benchmark', [Advances in Neural Information Processing Systems](https://arxiv.org/abs/2406.04391), Vol. 37, 2024, pp. 95266–95290.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y., 'Hellaswag: Can a machine really finish your sentence?', [arXiv preprint arXiv:1905.07830](https://arxiv.org/abs/1905.07830), 2019.

## List of abbreviations and definitions

**AGI** Artificial General Intelligence

**AI** Artificial Intelligence

**API** Application Programming Interface

**CoT** Chain of Thought

**FTE** Full-Time Equivalent

**FLOP** Floating Point Operations

**GPAI** General-Purpose Artificial Intelligence

**GPAI models** General-Purpose Artificial Intelligence models

**GPAISRs/GPAISR/GPAI+SR** General-Purpose Artificial Intelligence models with Systemic Risk

**LLM** Large Language Model

**PCA** Principal Component Analysis

**List of figures**

**Figure 1.** Illustration of the proposed computation of the principal capability measures. . . . . 13

**Figure 2.** Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025). . . . . 28

**Figure 3.** Robustness analysis. Noise on labels. Perturbations with Gaussaian noise with zero mean and standard deviation of 0.025, 0.05 and 0.1. Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025). . . . . 29

**Figure 4.** Robustness analysis. Removal of columns (one column at a time). Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025). . . . . 30

**List of tables**

**Table 1.** Results of the proposed approach for 33 models and 4 benchmarks, sorted by PC1 metric. . . . . 25

## Annexes

### Annex 1. Concrete example of benchmark aggregation

We provide a mock example to show the procedure of aggregating performance across benchmarks. The benchmarks are not exactly the same benchmarks as we suggest in the recommendation. We have collected the data by running four benchmarks on 33 models. Note that this is a proof-of-concept implementation and has many known failures, e.g. the elicitation for DeepSeek-R1 was broken. The purpose of this effort was to demonstrate that the approach is possible. A real implementation would have to be significantly more rigorous.

We have compiled the data in Table 1, which includes four benchmarks and 33 models. A small number of entries were filled in using PCA imputation because the data was not available. The table is sorted by performance according to PC1.

**Table 1:** Results of the proposed approach for 33 models and 4 benchmarks, sorted by PC1 metric.

<b>Model name</b>	<b>MMLU-Pro</b>	<b>GPQA-diamond</b>	<b>MATH-level-5</b>	<b>HumanEval</b>	<b>PC1</b>
openai/o3-mini-2025-01-31	0.792	0.662	0.952	0.970	3.107
openai/o1-preview	0.813	0.717	0.730	0.951	2.792
openai/o1-mini	0.739	0.596	0.880	0.939	2.525
google/gemini-1.5-pro-002	0.763	0.611	0.715	0.860	2.099
together/deepseek-ai/DeepSeek-V3	0.763	0.551	0.662	0.884	1.827
anthropic/claude-3-5-sonnet-20241022	0.775	0.561	0.566	0.939	1.779
google/gemini-1.5-flash-002	0.684	0.535	0.614	0.823	1.337
together/Qwen/Qwen2.5-72B-Instruct-Turbo	0.714	0.449	0.603	0.823	1.108
openai/gpt-4o	0.735	0.439	0.476	0.902	0.993
mistral/mistral-large-2411	0.697	0.419	0.556	0.866	0.942
openai/gpt-4o-mini	0.632	0.444	0.521	0.878	0.798
bedrock/amazon.nova-pro-v1.0	0.688	0.439	0.492	0.817	0.739
together/meta-llama/Meta-Llama-3.1-405B-Instru	0.732	0.359	0.452	0.829	0.523
anthropic/claude-3-opus-20240229	0.694	0.444	0.377	0.811	0.493
anthropic/claude-3-5-haiku-20241022	0.629	0.343	0.468	0.860	0.301
mistral/mistral-small-2501	0.642	0.338	0.468	0.774	0.147

<b>Model name</b>	<b>MMLU-Pro</b>	<b>GPQA-diamond</b>	<b>MATH-level-5</b>	<b>HumanEval</b>	<b>PC1</b>
bedrock/amazon.nova-lite-v1:0	0.596	0.384	0.380	0.835	0.096
together/meta-llama/Meta-Llama-3.1-70B-Instruc	0.682	0.343	0.368	0.744	-0.023
together/Qwen/Qwen2.5-7B-Instruct-Turbo	0.566	0.364	0.336	0.823	-0.173
together/meta-llama/Llama-3-70b-chat-hf	0.610	0.369	0.255	0.774	-0.326
together/google/gemma-2-27b-it	0.568	0.384	0.285	0.707	-0.450
google/gemini-1.0-pro	0.467	0.500	0.271	0.634	-0.510
anthropic/claude-3-haiku-20240307	0.493	0.298	0.442	0.726	-0.527
bedrock/amazon.nova-micro-v1:0	0.465	0.263	0.446	0.756	-0.646
together/meta-llama/Llama-3-8b-chat-hf	0.418	0.298	0.477	0.579	-0.934
anthropic/claude-3-sonnet-20240229	0.567	0.313	0.215	0.646	-0.969
mistral/ministral-8b-2410	0.430	0.278	0.134	0.726	-1.468
together/meta-llama/Llama-2-7b-chat-hf	0.083	0.393	0.733	0.152	-1.752
together/meta-llama/Meta-Llama-3.1-8B-Instruct	0.470	0.106	0.253	0.652	-1.800
together/google/gemma-2-9b-it	0.500	0.283	0.283	0.220	-1.930
together/deepseek-ai/DeepSeek-R1	0.499	0.071	0.265	0.226	-2.658
together/Qwen/QwQ-32B-Preview	0.027	0.015	0.024	0.805	-3.477
together/meta-llama/Llama-2-13b-chat-hf	0.000	0.393	0.000	0.000	-3.963

*Source: Own elaboration.*

It has to be noted that DeepSeek-R1 has a very bad score because it constantly runs out of tokens before submitting an answer. This makes it a special case. There might be more such special cases that we have not yet identified. This table has to be seen as an illustrative example of what it could look like, not as the final version.

For illustrative purposes (see Figure 2), as of February 2025, let's arbitrarily consider gpt-4o as the threshold for high-impact capabilities of GPAI models (as of July 2025 a more realistic capable model would be OpenAI o1 or Claude Sonnet 4).

In this example, the absolute normalised weights of the benchmark contributions (i.e. summing to 1) are:

- MMLU-Pro: 0.264
- GPQA-diamond: 0.251
- MATH-level-5: 0.251
- HumanEval: 0.233

This makes them almost uniform, which indicates that the PCA does not treat one of the datasets as significantly more informative than any other.

The first three principal components explain 0.69, 0.18, and 0.06 of the variance respectively. This is less than in the original observational scaling law paper. Given that the weights are almost uniformly distributed and all scores are between 0 and 1, a simple average of the scores would likely correlate heavily with PC1. However, since the scores could be on an arbitrary scale, we normalise each column before running the PCA. Thus, the final equation to compute PC1 values is:

$$PC1_{value} = \sum_i pc1_{component,i} \times \frac{x_i - mean_i}{scale_i}$$

However, for fixed mean,  $pc1_{component}$  and scale, this can be simplified. First, we have to pre-compute the following values:

$$weights = pc1_{component} / scale$$

$$bias = -mean \cdot weights = -mean \cdot pc1_{component} / scale$$

And then:

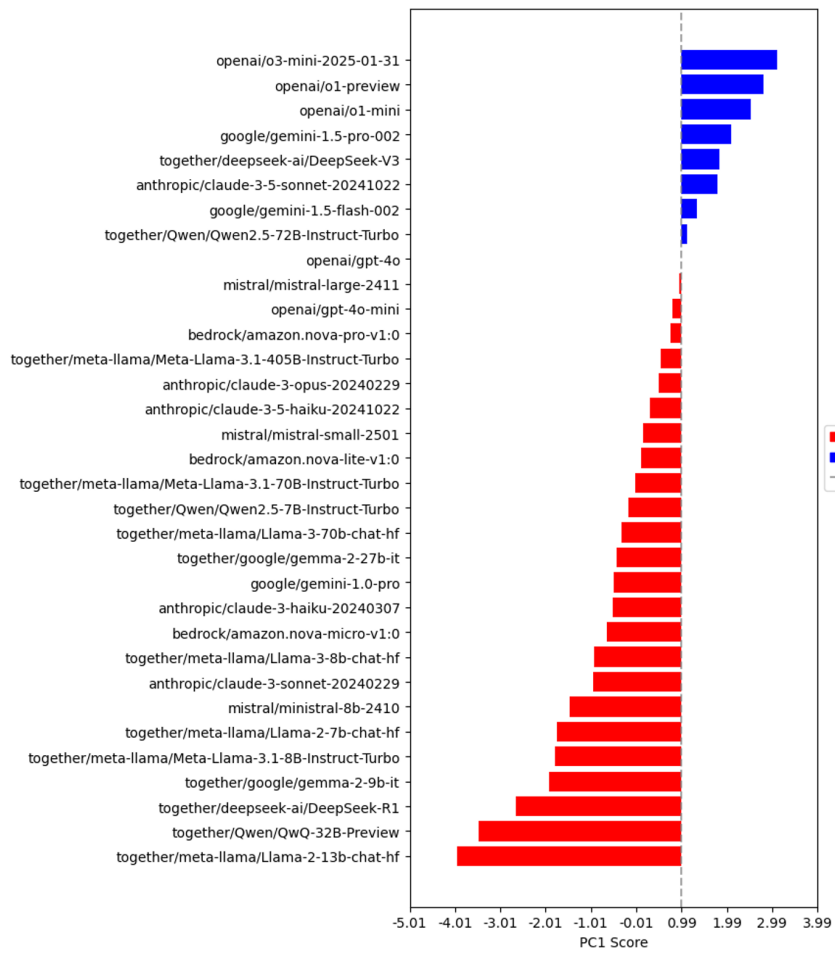
$$PC1_{value} = \sum_i x_i \times weights_i + bias$$

Which in our case means (numbers only displayed to 2 sig. digits for visibility):

$$PC1_{value} = 2.60 \times \text{MMLU-Pro} + 3.27 \times \text{GPQA-diamond} + 2.30 \times \text{MATH-level-5} + 1.98 \times \text{HumanEval} - 5.25$$

To compute the value for any new model, model providers can simply plug their numbers into the projection themselves.

**Figure 2:** Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025).



Source: Own elaboration.

## Annex 2. Robustness tests

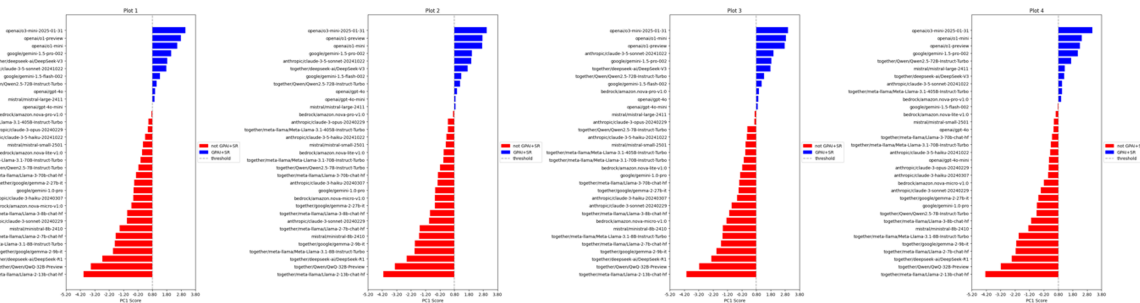
We test the robustness of the PC-1-based method for three different perturbations: a) add noise to scores, b) remove columns (i.e. benchmarks), c) remove rows (i.e. models). In all cases, we show how the score changes for each model and then compute the Kendall Tau value to give a sense of how much the ranking changed.

Note that this is a preliminary proof-of-concept robustness analysis and presents some known issues. The purpose of this effort was to demonstrate that the approach is sufficiently robust. A more comprehensive and rigorous analysis would be needed.

### Noise on labels

We add Gaussian noise with zero mean and standard deviation of 0.025, 0.05 and 0.1. Note that these are big perturbations and likely much bigger than even the differences between zero-shot and few-shot.

**Figure 3:** Robustness analysis. Noise on labels. Perturbations with Gaussian noise with zero mean and standard deviation of 0.025, 0.05 and 0.1. Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025).



Source: Own elaboration.

We find that the plots look pretty similar. Quantitatively, the Kendall Tau values are:

1. Std=0.025: 0.9659 (p-value = 0.0000)
2. Std=0.05: 0.9129 (p-value = 0.0000)
3. Std=0.1: 0.7208 0.8182 (p-value = 0.0000)

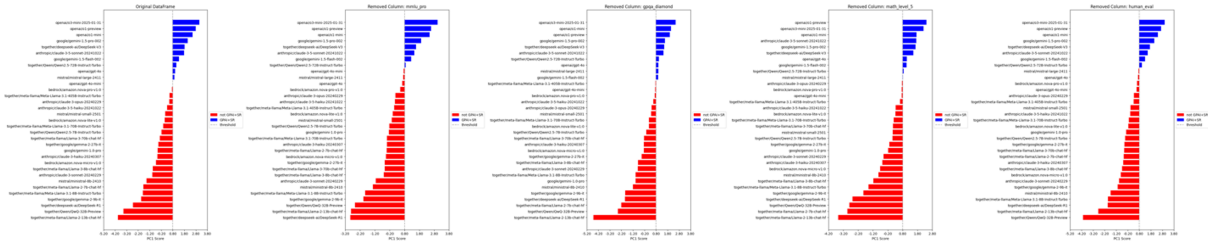
Which indicates only small changes in the rank order.

### Removal of columns (benchmarks)

We remove one column at a time and find that there are almost no perturbations in the rank order:

1. Column Removed: MMLU-Pro, Kendall Tau: 0.9129
2. Column Removed: GPQA-diamond, Kendall Tau: 0.9280
3. Column Removed: MATH-level-5, Kendall Tau: 0.9205
4. Column Removed: HumanEval, Kendall Tau: 0.9318

**Figure 4:** Robustness analysis. Removal of columns (one column at a time). Ranked models according to the PC1 value, and threshold on gpt-4o (as for February 2025).



Source: Own elaboration.

We remove combinations of columns as well such that we only have two remaining benchmarks respectively:

1. Columns Removed: (MMLU-Pro, GPQA-diamond), Kendall Tau: 0.8826
2. Columns Removed: (MMLU-Pro, MATH-level-5), Kendall Tau: 0.8674
3. Columns Removed: (MMLU-Pro, HumanEval), Kendall Tau: 0.7992
4. Columns Removed: (GPQA-diamond, MATH-level-5), Kendall Tau: 0.8485
5. Columns Removed: (GPQA-diamond, HumanEval), Kendall Tau: 0.8826
6. Columns Removed: (MATH-level-5, HumanEval), Kendall Tau: 0.8636

### Removal of rows (models)

We remove different numbers of rows at a time. We randomly draw a subset of N rows, remove them and compute the Kendall Tau value compared to the original ranking. We repeat this 100 times and average to reduce noise. We even test removing everything but 3 rows (i.e. removing 30 out of 33) and still get pretty stable Kendall Tau values.

We find the following values:

1. Mean Kendall Tau over 100 iterations (N=1): 0.9990
2. Mean Kendall Tau over 100 iterations (N=2): 0.9981
3. Mean Kendall Tau over 100 iterations (N=5): 0.9968
4. Mean Kendall Tau over 100 iterations (N=10): 0.9929
5. Mean Kendall Tau over 100 iterations (N=20): 0.9777
6. Mean Kendall Tau over 100 iterations (N=30): 0.8133

## Getting in touch with the EU

### In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: [european-union.europa.eu/contact-eu/write-us\\_en](https://european-union.europa.eu/contact-eu/write-us_en).

## Finding information about the EU

### Online

Information about the European Union in all the official languages of the EU is available on the Europa website ([european-union.europa.eu](https://european-union.europa.eu)).

### EU publications

You can view or order EU publications at [op.europa.eu/en/publications](https://op.europa.eu/en/publications). Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex ([eur-lex.europa.eu](https://eur-lex.europa.eu)).

### EU open data

The portal [data.europa.eu](https://data.europa.eu) provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

# Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



Scan the QR code to visit:

**[The Joint Research Centre: EU Science Hub](https://joint-research-centre.ec.europa.eu)**

<https://joint-research-centre.ec.europa.eu>



Publications Office  
of the European Union